MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY

and

CENTER FOR BIOLOGICAL INFORMATION PROCESSING
WHITAKER COLLEGE

A.I. Memo 770
C.B.I.P. Paper 003

January, 1984

# Selecting One Among the Many: A Simple Network Implementing Shifts in Selective Visual Attention

**Christof Koch and Shimon Ullman**

This study addresses the question of how simple networks can account for a variety of phenomena associated with the shift of a specialized processing focus across the visual scene. We address in particular aspects of the dichotomy between the preattentive-parallel and the attentive-serial modes of visual perception and their hypothetical neuronal implementations. Specifically, we propose the following:
(1) A number of elementary features, such as color, orientation, direction of movement, disparity etc. are represented in parallel in different topographical maps, called the early representation.
(2) There exists a selective mapping from this early representation into a more central representation, such that at any instant the central representation contains the properties of only a single location in the visual scene, the *selected* location.
(3) We discuss some selection rules that determine which location will be mapped into the central representation. The major rule, using the saliency or conspicuity of locations in the early representation, is implemented using a so-called Winner-Take-All network. A hierarchical pyramid-like architecture is proposed for this network. We suggest possible implementations in neuronal hardware, including a possible role for the extensive back-projection from the cortex to the LGN.



This report describes research done within the Artificial Intelligence Laboratory and the Center for Biological Information Processing (Whitaker College) at the Massachusetts Institute of Technology. Support for the A. I. Laboratory's research in artificial intelligence is provided in part by the Advanced Research Projects Agency of the Department of Defense under Office of Naval Research contract N00014-80-C-0505. The Center's support is provided in part by the Sloan Foundation and in part by Whitaker College.

## 1. Introduction

A number of psychophysical studies concerning the detection, localization and inspection of objects in the visual field have suggested a two-stage theory of human visual perception. The first stage is the "preattentive" mode, in which simple features are processed rapidly and in parallel over the entire visual field. In the second, "attentive" mode, a specialized processing focus, usually called the focus of attention, is directed to particular locations in the visual field. The analysis of complex forms and the recognition of objects are associated with this second stage (Neisser, 1967; Bergen & Julesz, 1983; Treisman, 1983; Ullman, 1983; Julesz, 1984). The computational justification for such a hypothesis comes from the realization that while it is possible to imagine specific algorithms performing specific tasks such as inspection, counting, marking etc. at specific locations, it is difficult to imagine these algorithms operating in parallel over the whole visual scene, since such an approach will quickly lead to a combinatorial explosion in terms of required computational resources (Ullman, 1983; Poggio, 1984). This is essentially the major critique of Minsky and Papert to a universal application of perceptrons in visual perception (Minsky & Papert, 1969). Taken together, these empirical and theoretical studies suggest that beyond a certain preprocessing stage, the analysis of visual information proceeds in a sequence of operations, each one applied to a selected location (or locations).

The sequential application of operations to selected locations raises two central problems. First, what are the operations that the visual system can apply to the selected locations? Second, how does the selection proceed? That is, what determines the next location to be processed, and how does the processing shift from the current to the next selected location? In this paper we consider primarily the second of these questions. With respect to the first question, we only suggest that one of the fundamental operations is what we term *"selective mapping"*. According to this view, the early "preattentive" representations describe the visual scene in terms of a number of elementary properties such as color, orientation, depth and movement (Treisman, 1983; Julesz & Bergen, 1983). When a location is selected, its properties are mapped from the early representations into a higher, central representation. This central representation consequently contains the properties of only this single selected location. With respect to the second question, we discuss a number of "selection rules" that determine the next location to be processed. It is obviously desired that the selective mapping does not occur at random, but is applied to "interesting" locations. But how does one define what "interesting" means, without having to recourse to higher, symbolic concepts? We will propose a specific set of "selection rules" that determine which location will be mapped into the central representation at any given time. The major rule for the initial selection of a location is based on the conspicuity of that location, i.e. by how much its properties differ from the property of its neighborhood. Two rules for shifting from one

selected location to another are based on (i) proximity and (ii) similarity with the presently selected location. We will propose and discuss simple neuron-like networks that implement the selective mapping and the selection rules. Formulating the mechanism related to selective visual attention in terms of a schematic network where the individual components perform simple local operations, rather than in the language of higher cognitive concepts, has the advantage that specific predictions concerning the anatomy and electrophysiology of the specialized cortical regions involved in attention can be derived. The main point we wisk to make is not so much that the particular network we propose is necessarily implemented in the brain, but that the shift of selective visual attention and related visual operations can be explained and modeled using simple mechanisms compatible with cortical physiology and anatomy.

## 2. The Early Representation

According to our suggestion, selective visual attention operates on what we call the *early representation*, a set of topographical, cortical maps encoding the visual environment (Barlow, 1981). The early representation includes a variety of different maps for different *elementary features* such as orientation of edges, color, disparity or direction of movement (see figure 1). Neighborhood relations are preserved in these maps, i.e. nearby locations in the visual scene project to nearby locations in the map. Local, inhibitory connections, mediating lateral inhibition, occur either at an earlier stage or within the feature maps. Thus, locations that differ significantly from their surrounding locations are singled out at this level. The state of each of these maps signals how conspicuous a given location in the visual scene is: a red blob surrounded by similar red blobs will certainly be less conspicuous than a red blob surrounded by green blobs. It should be emphasized that the different maps do not necessarily have to be in physically different locations, but may be intermixed. Moreover, these maps may possibly exist at different scales, i.e. at different spatial resolutions, in accordance with the evidence for multiple spatial channels (e.g. Wilson & Bergen, 1979).

In addition to the maps for the different features, we assume the existence of another topographical map, termed the *saliency map*, which combines the information of each individual map into one global measure of conspicuity. Points in the elementary feature maps, corresponding to one location in the visual scene, project onto a unit in the saliency map. The saliency map gives a "biased" view of the visual environment, emphasizing interesting or conspicuous locations in the visual field. Since the saliency map is still a part of the early visual system, it most likely encodes the conspicuity of objects in terms of simple properties such as color, direction of motion and orientation. Saliency at a given

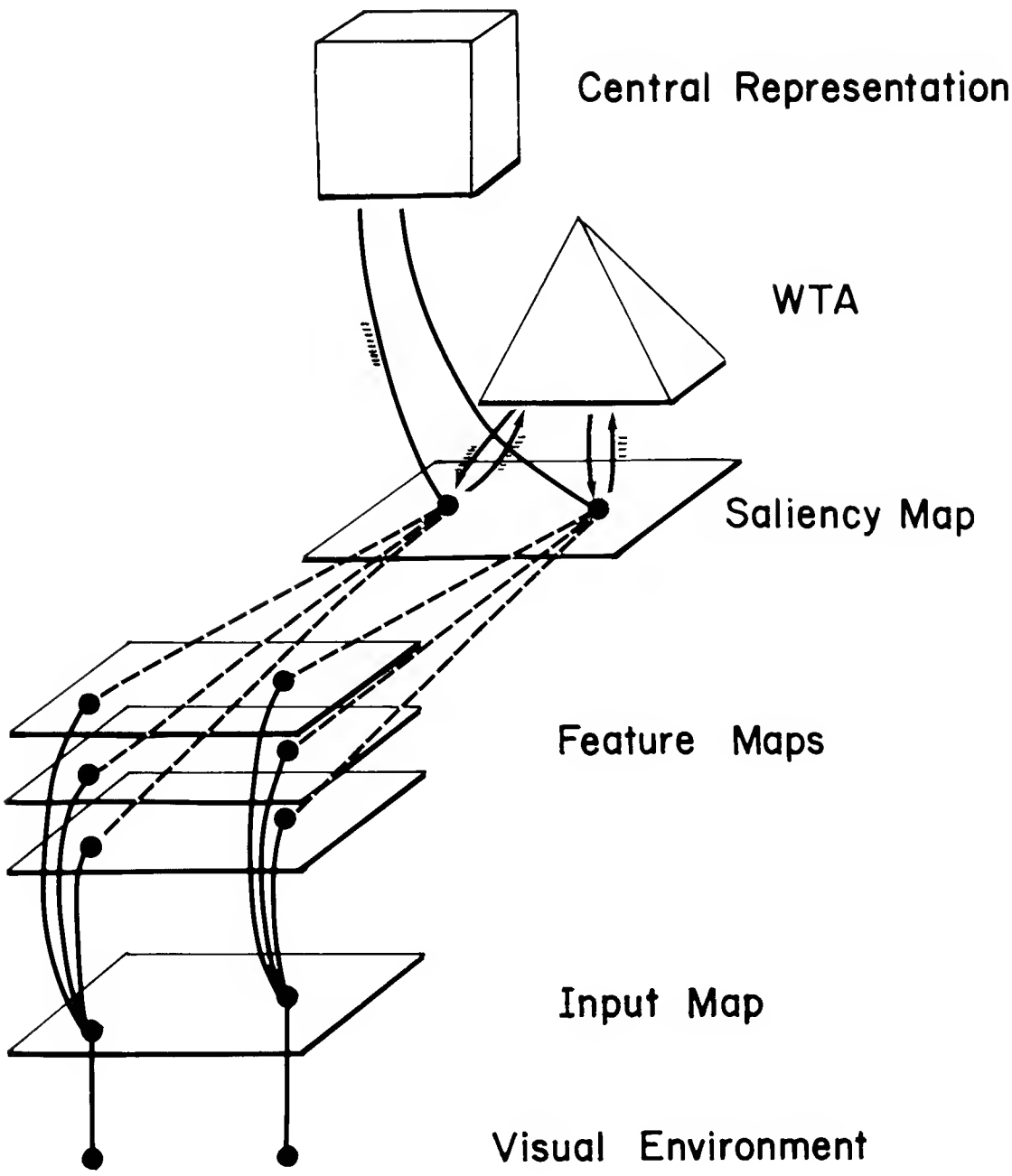


**Figure 1.** A highly schematic drawing illustrating the main elements of the selection process. Every location in the visual environment is analyzed and segmented in terms of elementary features, such as color, orientation, disparity etc. and represented in the corresponding feature map. The information in these different maps is combined in the saliency map, encoding conspicuous locations in the visual scene. The WTA network subsequently routes the properties of the most conspicuous location to the central representation. After the selection process, the central representation contains the properties of a single, the selected, location.

location is determined primarily by how different this location is from its surround in color, orientation, motion, etc. It is possible, however, that the relative weight of the different properties contributing to this representation can be modulated by the activity of some higher cortical centers.

One general problem that must be solved at this level is aligning the feature maps with respect to each other, i.e. solving the *spatial register* problem. Combining the information

of the different feature maps or retrieving information relevant to a single location, requires a fast and reliable pathway to address the same location in the different maps. We suggest that this register is obtained as part of the selective mapping process, described next.

## 3. Selective Mapping

We assume that in addition to the early representation, the properties of objects can also be represented in a second, "central" representation. With respect to the connections between the early and the central representation, we make the simple assumption that all the units in one feature map project to a single unit (not necessarily a single cell) in the central representation. For example, all units that detect the presence of a vertical orientation anywhere in the visual field are connected to a central "verticality detector". If the central unit is active, it can be inferred that there is at least one object with the specific feature somewhere present in the visual field. The mapping from the individual feature map to the feature detector cell will not preserve the location. Retrieving location, except perhaps in a rough manner (e.g. there is a red object somewhere in the lower left hemisphere), cannot be done at this parallel stage. The central representation thus signals the presence of at least one instance of a given property, such as being red or horizontal. Without the use of selective attention it has no way of "knowing", however, whether different attributes belong to a single object (which is both red and horizontal), or to different objects in the visual field. Finally, if the projection from the early maps to the feature detector can distinguish between one, two, three and four or more active units, it would constitute a sort of parallel counting device, alleviating the need for any perceptron-like network of the kind explored by Minsky and Papert (1969) to compute the predicate: "the visual scene contains exactly $i$ points". Such a mechanism may explain the ease and proficiency with which human observers can count the number of up to four or five objects present in the visual field (Atkinson, Campbell & Francis 1976).

Consider next the problem of performing a conjunctive search task (Treisman & Gelade, 1980), for instance, searching for a target line segment that is both vertical and green in the presence of red and green lines that are either horizontal or vertical. The mechanisms considered so far are insufficient for this task: the central representation will signal the presence of red, green, vertical, and horizontal objects, but will not represent explicitly the combination of these properties. For computing conjunctions we postulate a "switch" that routes the properties of a single location, the *selected* or *attended* location, into the central representation, which will now only contain information relevant to the selected location. Note that the computations required to abstract properties from the information contained in the visual input map are performed within the early representation, i.e. prior to the

selection process, and not subsequent to it. This distinction is important, for example, in the computation of color. As has been demonstrated psychophysically (e.g. Land, 1959), the computation underlying color perception is a global process, that requires the entire visual field (or a large portion of it). It is therefore reasonable to assume that the computation of color and other properties proceeds within the early representation, prior to the selection of a location for further processing.

### 3.1 Selective Mapping: The Basic Mechanism

The operations underlying the selective routing of information from the early representation to the central one can be performed by two complementary cellular networks (see also figure 5). One such network, called the *Winner-Take-All* network (WTA network; see Feldman, 1982, who introduced this term, Feldman & Ballard, 1982) localizes the most active unit in the saliency map while the second network relays the properties of the selected location to the central representation. It is here that specialized visual routines for the extraction of shape and form can be applied to the properties of the selected location. At any given time only one location is selected from the early representation and copied into the central representation. The WTA network, equivalent to a maximum finding operator, operates on the output $x_i$ of the units in the saliency map. In a neural network $x_i$ can be interpreted as the electrical activity (intracellular voltage or spiking rate) of the unit at location $i$. The WTA mechanism maps this set of input units onto an equal number of output units, described by $y_i$, using the transformation rule:

$$\begin{aligned} y_i &= 0 \quad if \quad x_i < \max_j\{x_j\} \\ y_i &= f(x_i) \quad if \quad x_i = \max_j\{x_j\} \end{aligned} \tag{1}$$

where $f$ is any increasing function of $x_i$ (including a constant). All output units are set to zero except the one corresponding to the most active input unit.

### 3.2 The Winner-Take-All Network

Building a WTA network may appear as a straightforward task, but complications arise when the intrinsic properties of biological hardware are taken into account. Depending on the underlying hardware, two extremes for computing the maximum of a given set can be envisioned. On a serial machine, the simplest algorithm is a sequential search for the largest number through the entire input set. The drawback to this method is that for $n$ discrete inputs, $n$ basic time steps are required (by basic time step we always refer to the time required to execute an elementary operation such as comparing two numbers). A

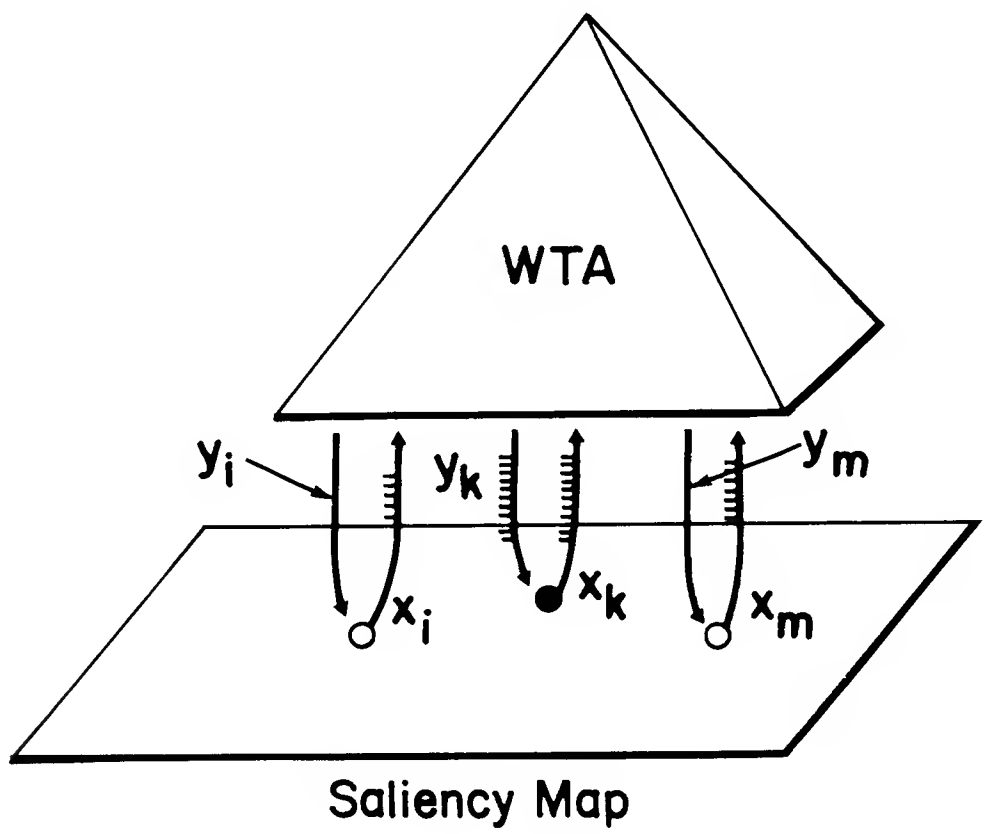


**Figure 2.** A schematic drawing illustrating the Winner-Take-All (WTA) network computing the maximum $x_k$ of a set of $n$ input units in the salient map. It localizes the most conspicuous point by a number of parallel operations and activates the corresponding output line (in this case unit $x_k$) after at most $2log_m n$ time steps (if $m$ units can be compared simultaneously).

highly parallel machine with $n$ processors, each one having direct access to the other $n-1$ processors, can compute the maximum in one time step by comparing simultaneously the value of each processor with the values of all the other processors.[1]

A simple implementation one may suggest for the WTA network is a *mutual inhibitory network* of the type studied by Hadeler (1974), where every unit inhibits every other unit. In these networks, neurons are assumed to be linear summation devices, followed by a threshold operation (see for instance McCullouch & Pitts, 1943). They can be described by

$$x_i^{t+1} = f(I_i^t + \sum_{j \neq i} w_{ij} x_j^t), \tag{2}$$

where $x_i^{t+1}$ is the output of unit $i$ at time $t+1$, $w_{ij}$ is the synaptic weight between the $i$-th and the $j$-th cell, $I_i$ is the input to the $i$-th cell and $f(x)$ is zero for all $x < x_{thres}$ and a monotonic increasing function of $x$ above this positive threshold value. Such a mutual inhibitory network will be unable, however, to implement the WTA computation. The reason is the following. The requirement that for any set of inputs $I_i$, only a single output survives implies $w_{ij} < -1$ for every $i, j$. If this condition is met, then for many input sets the network described by equation (2) will oscillate and fail to converge. Convergence is only guaranteed

[1]This is essentially the mechanism Feldman and Ballard (1982) propose for their implementation of a WTA network.

if the largest input is larger than the sum of the other inputs, $I_i > \sum I_j$. Otherwise, the network will oscillate. Thus, in practice, these networks will fail for more than 2 units. We conjecture that this undesirable property is unavoidable in any cellular network where the individual components have only access to the summed activity of the converging cells, i.e. $\sum w_{ij} x_j^t$. A possible remedy for this problem is the introduction of an amplitude-dependent time-course of $x_i^t$. If, for instance, the unit with the largest output responded faster than units with smaller output, it could inhibit its competitors before it would be inhibited by them, thus avoiding oscillations. In this manner it is possible to combine more that two units, but it would still be difficult to construct a network of this type with large number of elements.

In the previous discussion neurons were assumed to be simple, linear threshold devices. It has been, however, long realized that neurons are complicated computational machines performing a variety of logical operations on their input (e.g. Schmitt, Dev & Smith, 1976). Even taking account of these more realistic neurons, it is still difficult to envisage an implementation of the WTA computation by a single uniform network. Moreover, the requirement that each unit in the network is connected to every other unit seems prohibitive in terms of numbers and the non-locality of the connections.

We therefore propose a different cellular mechanism, based on the following two assumptions.

(1) Except for some long-range excitatory connections, most connections, whether excitatory or inhibitory, are local.
(2) Each elementary processing unit only performs some simple well-specified operation, such as addition or multiplication. In particular, the basic processing units are unable to use any symbolic information, such as addresses.

The basic version of the WTA network consists of two intercalated pyramid-like structures.[2] The network operates in a highly parallel fashion by computing the maximum of a small number $m$ of units across the whole input set. Next, comparisons are made among these local maxima to compute again the most active unit. These comparisons are repeated $k = log_m n$ number of times until the global maximum has been determined. Figure 3 shows one particular implementation of the WTA network with $m = 2$. The more active unit inhibits the less active unit and transmits its activity onto the next higher level. Here, among $n/2$ units, the process is repeated.[3] Under the assumption that the connections between the levels transmit faithfully the activity of the units, the top-most unit in the pyramid will hold the activity $x_i$ of the global maximum after $k$ time steps. However, it is the location of the

[2]Hierarchical, pyramid-like computer architectures have been proposed for image processing and analysis. For an overview of their use see (Rosenfeld, 1984).

[3]The computational structure is similar to the Wimbledon tennis tournament where players drop out if they lose a single match (a so-called knock-out competition).

maximum and not its absolute value which is of relevance for the selection process. The location of the corresponding unit in the saliency map can be obtained by the use of the second pyramid, having a reversed flow of information with respect to the first pyramid. It "marks" the path of the most active unit through the first pyramid, activating finally the output $y_i$ of the WTA. This is done with the help of an *auxiliary* unit associated with every unit in the first pyramid (called the *main* unit). The auxiliary unit is only activated if it receives conjoint excitation from its main unit and from the auxiliary unit at the next higher level. Since at every level the most active (main) unit in a local comparison suppresses the activity of the other $m-1$ (main) units, the associated auxiliary units as well as all auxiliary units in the subtree below them can never be activated. After another $k$ time steps, the output $y_i$, corresponding to the most active unit in the saliency map, will be activated, while the rest of the output units remain silent. Except for the pathological case when two or more inputs are exactly equal, the WTA network will always converge to a unique solution within at most $2log_m n$ time steps. It can be built with no more than $2n\frac{m}{m-1}$ units. This is immediately established by adding up the total number of units at each of the $k = log_m n$ levels

$$n + \frac{n}{m} + \frac{n}{m^2} + \cdots + \frac{n}{m^k}.$$

This expression is smaller than the infinite geometric series

$$n(1 + m^{-1} + m^{-2} + \cdots + m^{-k} + m^{-k-1} + \cdots),$$

since the terms of the order of $m^{-(k+1)}$ and higher vanish. This series converges to $n\frac{m}{m-1}$, if $m > 1$. The factor 2 takes both pyramids into account. Notice, that for all integers $m$, the WTA network can always be built with less than $4n$ units. Assuming that the optic nerve contains approximately $10^6$ fibers and that $m = 10$ neurons can compare their activity simultaneously, a WTA network covering the entire retinal image would require no more than $2.2 \cdot 10^6$ neurons, a small fraction of all visual neurons. If only the Y-system, with its associated short delay, high movement sensitivity, large receptive fields and transient temporal response, provided the major input to the WTA network, this number would drop substantially. In the cat, about 4% of all ganglion cells are of the Y-type. If this percentage carries over to primates and man, a WTA network for the entire visual field could be built with just $10^5$ neurons. Interestingly, the computational architecture of the WTA network is reminiscent of the K- and P-pyramids proposed by Minsky for his K-line theory of memory (Minsky, 1979).

### 3.3 Mapping the Selected Location into the Central Representation

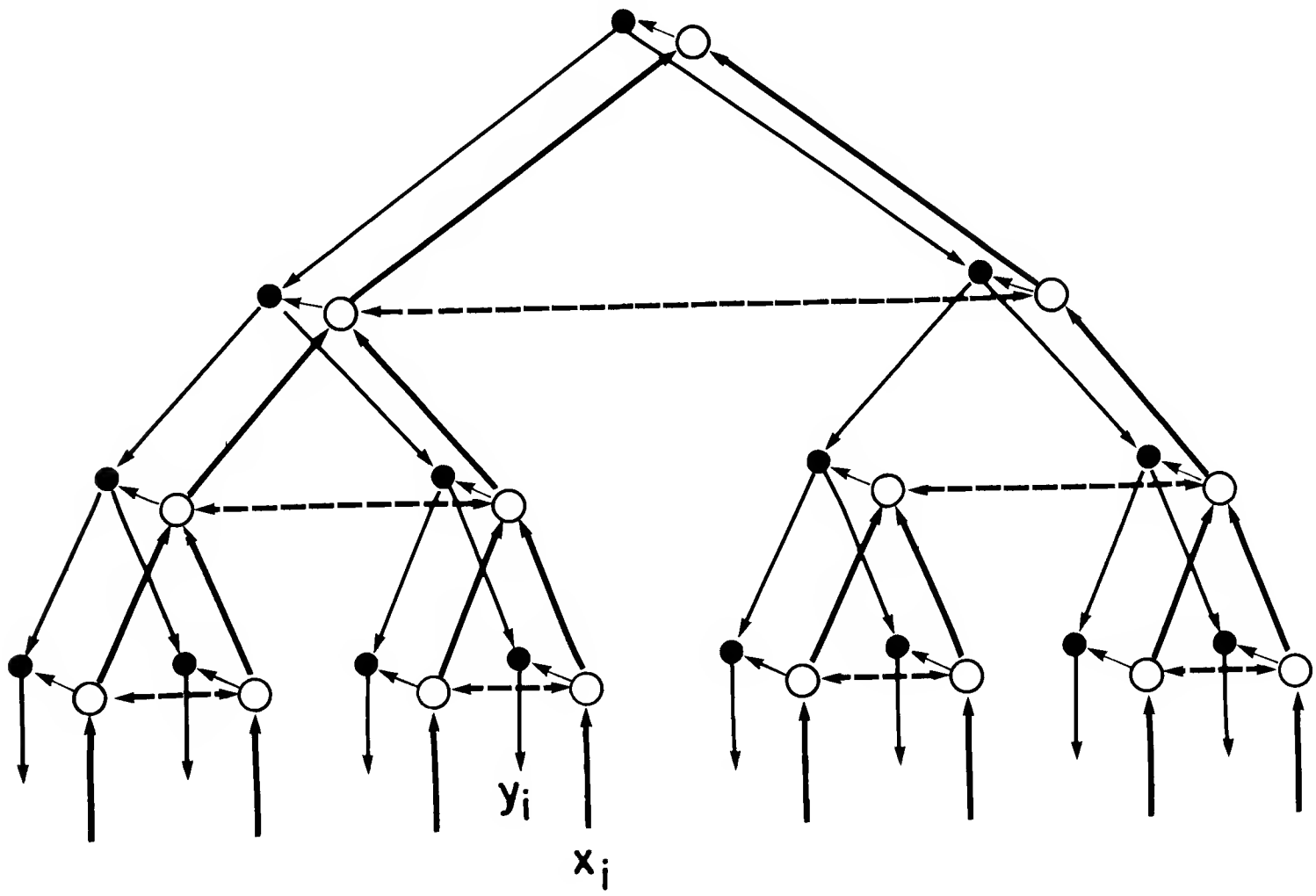


**Figure 3.** A possible implementation of the Winner-Take-All network with $n = 8$ input units. The local comparison takes place between $m = 2$ units. The more active unit inhibits the less active one and excites the unit on the next level. The auxiliary units, drawn in black, are only activated if they receive conjointly excitation from their associated main unit and from the auxiliary unit at the higher level. The auxiliary unit $y_i$, corresponding to the most active unit $x_i$ in the saliency map, will be activated after at most $2log_m n = 8$ time steps. In order to insure stability against noise and to enforce neighborhood relations between all neighboring points (for instance between the two middle units, belonging to two different subtrees) additional connections (and units) can be added between (and within) levels. We have just shown the most sparse implementation of a WTA network.

Once the most conspicuous point has been localized in the saliency map, its properties, i.e. the information contained within the early representation, must be copied into the central representation. The routing of this information can be achieved by removing some tonic inhibitory influence or by increasing the amount of excitation at the selected location in the early representation. We will not suggest here specific mechanisms for the mapping operation. The crucial point is that the WTA network directs the "copy" operation to a single selected location (figure 2). Note, that the selection system itself is *not* responsible for the information processing relevant to the visual task but simply selects which area of visual space should be inspected (Posner, Snyder & Davidson, 1980). It can be likened to a spotlight illuminating some portion of the visual field. This view of attention is in accordance with the fact that it is not possible for visual attention to be allocated simultaneously to two different positions in space (Posner *et al.*, 1980).

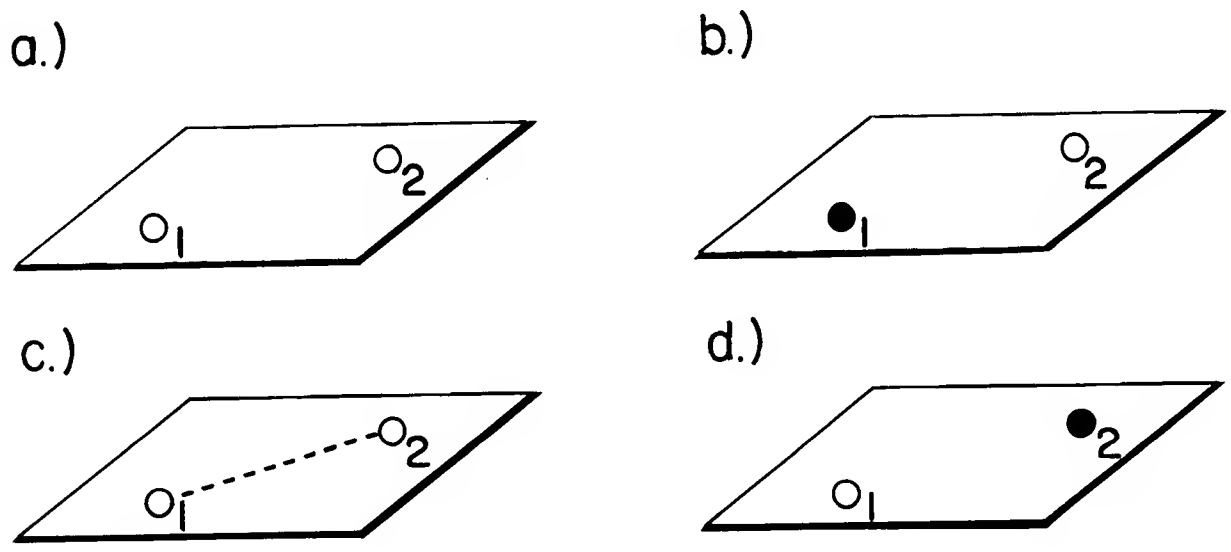


**Figure 4.** Shifting visual attention within the salient map. Once the most conspicuous location (point 1) has been detected and examined, its corresponding output $x_q 1$ decays and the WTA mechanism shifts to the next most salient location, 2. The time needed to find the next location increases with increasing distance between locations 1 and 2.

### 3.4 Shifting the Processing Focus

Until now we have only considered the initial selection of an "interesting" location. But how does the selection process move from one location to the next, i.e. how can selective attention shift across the visual field (Shulman, Remington & McLean, 1979)? From psychophysical experiments it is known that it takes some measurable time to shift the focus of attention from one location to another (Eriksen & Schultz, 1977; Tsal, 1983). There is some evidence that this time increases with the distance between these locations (Shulman *et al.* 1979; Tsal 1983; see, however, Remington & Pierce, 1984).

A simple way to introduce such dynamics into our model is to let the conspicuity of the maximal active unit in the saliency map decay, even if constant visual stimuli are present. This decay may be implemented either locally or centrally (or by some combination of the two methods). By "local" we mean that an active location in the saliency map adapts and decays after a while. By "central" we mean that once the information from the early representation has been relayed to the central representation a signal is sent back, inhibiting the most active unit in the saliency map, i.e. its conspicuity fades. The WTA network responds to the new input configuration by shifting away from the presently selected location and towards the next most conspicuous location. The convergence time, i.e. the time taken by the WTA network to converge to the newly selected location, depends primarily on the distance between the two locations. In the worst case it will take $2log_m n$ time steps for the new maximum to propagate up, and subsequently down, the $log_m n$ layers (see figure 4), assuming that the comparison of $m$ units can be done in one time step.

Shorter convergence time can be achieved if the two locations are close to each other. Note, that the dependency of the convergence time on distance follows naturally from the computational architecture of the WTA network and does not have to be artificially imposed. In our previous example of a WTA network with $n = 10^6$ and $m = 10$, a solution will always be found after at most 12 time steps. Since time-constants for neurons are in the msec range, this number seems broadly compatible with the estimated $30 - 50$ msec required to shift visual attention to a new location (Bergen & Julesz, 1983). After a new location has been selected, the visual information associated with this location is routed to the central representation. The local scheme is similar, except that the most active unit is locally inhibited, for instance at some fixed time after the WTA mechanism has converged. These schemes are non-exclusive; in fact, it seems likely that some local, automatic mechanism might always be in operation. The cortical mechanism is only invoked when a voluntary shift of attention is desired (Posner, 1980). The basis for both mechanisms is a long-lasting inhibition of the selected unit in the saliency map preventing, for a given time period, that the attentional focus will revisit this location. A temporary inhibition, lasting more than 500 ms, has been reported by Posner, Cohen and Rafal (1982) after attentional shifts away from a cued location. Processing efficiency appears to be reduced from locations in the visual field once attention is withdrawn.

In summary, selective attention in the case we have considered requires three different stages (see figure 1). First a set of elementary features is computed in parallel across the visual field and is represented in a set of cortical, topographical maps. Locations in visual space that differ from their surround with respect to an elementary feature such as orientation, color or motion are singled out in the corresponding map. These maps are combined into the saliency map, encoding the relative conspicuity of the visual scene. Second, the WTA mechanism, operating on this map, singles out the most conspicuous location. Thirdly, the properties of this selected location are routed to the central representation. The WTA network then shifts automatically to the next most conspicuous location. The visual system processes a scene in a sequential and automatic way by selectively inspecting the information present in conspicuous locations. The mechanism sketched here might of course not only be used for the shift of the attentional focus but also for such visual routines as tracking of contours, counting objects or marking a specific location (Ullman, 1983).

## 4. Two Rules for Shifting the Processing Focus

Should there be any systematic relationship between the current location and the next location to be selected? If no such relationship is enforced, it would seem difficult to

visually inspect areas of the visual field without constantly shifting to conspicuous, but distant, locations. Objects tend to occupy a compact region in space with similar properties (color, motion, etc.). If the shifting apparatus is to scan automatically different parts of a given object, it is useful to introduce a bias based on both spatial proximity and similarity. Searching for an "interesting" target around a selected location would profit from a selection mechanism biased to nearby locations (what we call *proximity preference*). Scanning the visual field for objects with a common identifying feature, for instance the color red, would be likewise facilitated if locations with similar features to the presently selected location are preferrentialy selected (*similarity preference*). Both mechanisms are related to phenomena on perceptual grouping and "Gestalt effects" which occur as a function of object similarity and spatial proximity (Wertheimer, 1923; Beck, 1967). The next two section discuss these rules in more detail.

### 4.1 Proximity Preference

It would seem advantageous from a computational point of view, if the selection process shifts preferentially to conspicuous locations in the neighborhood of the presently selected location, instead of shifting to the global maximum independent of any locality considerations. Inspecting, for instance, part of a visual image for the occurance of some special feature (or conjunction of features) could be performed much more efficiently if the search is automatically limited to some neighborhood. The simplest way of implementing such a proximity preference within the framework of the WTA mechanism is to enhance all units in the neighborhood of the currently selected unit in the saliency map. Such a preference can be incorporated in a straightforward manner into the network described earlier. More specifically, we assume that the output of the WTA mechanism associated with the presently attended location enhances the conspicuity of nearby units in the saliency map by a factor depending on the distance between the location and its neighbors, thereby facilitating shifts of the processing focus to nearby locations. This is equivalent to postulating the existence of an attractive potential around every selected location. Some experimental evidence for this type of interaction is provided by Engel (1971, 1974). His results indicated that the probability of detecting a target depends on the proximity of the location being attended to.

### 4.2 Similarity Preference

On similar computational grounds one can justify the existence of a similarity preference. We postulate therefore the existence of an interaction between similar, elementary features:

the processing focus will preferentially shift to a location with the same or similar elementary features as the presently selected location. Such a mechanism assumes interactions within individual elementary feature maps, but not between them, and therefore it does not require precise topographic mappings between the different elementary feature maps. The interaction will be activated by the output of the WTA network. This output ($y_k$ in figure 2) increases the excitability, viz. the conspicuity, of all units in a neighborhood of the selected location within those elementary feature maps where the corresponding features have been detected. If the currently selected location contains for instance a red, horizontal line, then neighboring units in the feature map for horizontal and red will be facilitated. The processing focus will now preferentially shift to either red and/or horizontal targets. The effect of the similarity preference is opposite to the initial bias towards conspicuous locations. Locations with similar properties initially inhibit each other. After a location has been selected, it tends to facilitate the conspicuity of nearby locations with similar properties. Although the two processes have opposite effects, they can both be implemented without causing undesirable contradiction or interference. The first occurs early on within the individual maps and is implemented by local inhibition within the maps. The similarity preference can be implemented by a feedback from the output of the WTA network to all the different feature maps. Finally, it would be expedient if the similarity preference for individual features could be switched on or off voluntarily (look for red objects i.e. facilitate the red feature map), but it is unclear to what degree such a control actually exists.

A partial experimental support for this type of interactions comes from a recent study by Geiger and Lettvin (1984) who investigate the influence of the attended location on lateral masking. If the subject fixates a central point, while a group of three letters is flashed onto the screen at some distance from the central point, the subjects are usually unable to name the central letter. However, if a copy of the interior letter is flashed at the fixation point, the letter in the periphery transiently stands out against its neighbors in the string.

## 5. Biological Considerations

What could the anatomical correlate for our selection mechanism be? The maps for the different elementary features are most likely localized in areas within and beyond the striate cortex, such as MT and MST for motion, and perhaps V4 for color. Does it then necessarily follow that the saliency map, which combines aspects from the different elementary feature maps, must be located beyond these areas? One intriguing possibility is that the saliency map resides either at the level of the lateral geniculate nucleus (LGN) or in the striate cortex, V1 (see figure 5). The LGN in the cat and striate cortex in primates, represent the last major station along the retino-geniculo-cortico pathway before the visual information is

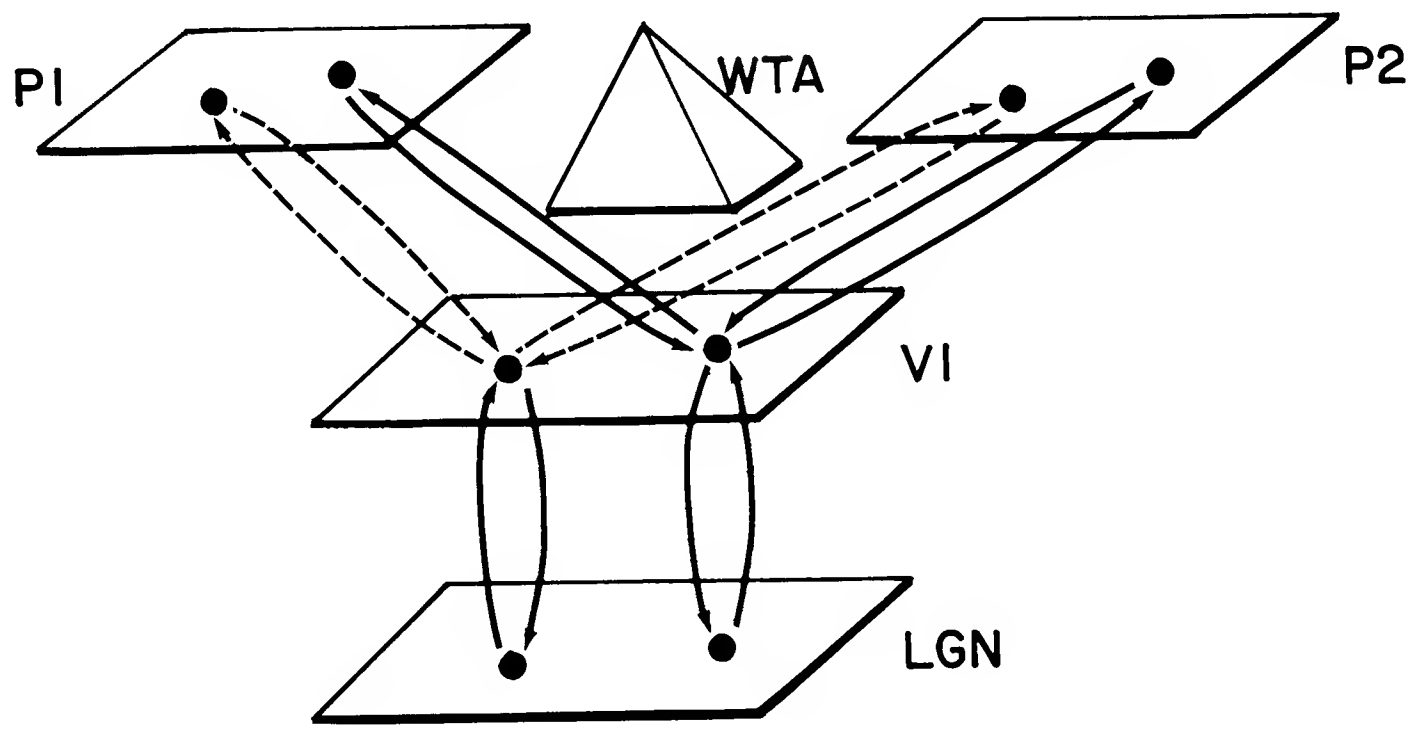


**Figure 5.** A biological implementation for the selection process. The saliency map may be localized either within the lateral geniculate nucleus (LGN) or within the striate cortex (V1). The backprojection from the different cortical maps for different properties (for instance P1 and P2) solve the spatial register problem. The WTA network selects the most active unit in the saliency map, subsequently routing the information corresponding to this *selected* location into the central representation. Interestingly, Crick proposed recently (1984; see also Yingling & Skinner, 1977) that the attentional searchlight is controlled by the thalamic reticular nucleus, a layered structure surrounding the thalamus. It receives extensive feedback from the visual cortex and projects onto the principal relay cells in the LGN.

dispersed to different regions. The Y-pathway projects to the striate and extrastriate areas V1, V2 and V3 in cat, but predominately to V1 in the monkey. The X-pathway behaves similar in both cat and monkey, projecting predominantly to V1. The W-pathway, much less explored, in addition to sending fibers to V1, V2 and V3 probably also innervates the medial Clare-Bishop area and area 21a (Graybiel & Berson, 1981; Sherman, 1984). One puzzling feature about the LGN is the existence of an extensive reciprocal projection from the cortex onto the LGN (Macchi & Rinvik, 1976). This connection observes the general principle that for every geniculocortical projection there is a corresponding corticothalamic pathway. Although little information on the number of fibers involved in this back projection is available, estimates suggest that it is at least as massive as the forward projection, and perhaps considerably stronger (Gilbert and Kelly (1975) estimate that about half of all cells in layer VI in the cat striate cortex send their axons to the LGN).

These strong reciprocal connections could be used to solve the spatial register problem in the manner suggested in figure 5. The visual environment is encoded at the level of the LGN or V1 in neurons having circular-symmetric receptive fields. Subsequently, different properties such as color, motion, disparity etc. are processed, analyzed and represented in

different regions of the cortex. These regions then project back to the LGN (via V1). If, for instance, in the area computing color a single location stands out among all others, this location will enhance the corresponding location in the LGN. Similarly, the different visual maps all feed back into the saliency map, providing it with a measure of the strength and importance of the different features. The WTA network now finds the most active unit in the saliency map.

This arrangement provides a mechanism for spatial register, since all the information pertaining to the selected location is transmitted together to the central representation. A notable limitation of this mechanism is that spatial register is obtained for one location at a time, a property that is consistent with psychophysical evidence (Treisman & Gelade, 1980). In terms of connections among different visual areas, this arrangement has two interesting properties. First, it requires an extensive topographic projection from the cortex back to the LGN (Tsumoto, Creutzfeldt & Legendy, 1978). Second, it does not require precise topographic reciprocal interconnections among all the different visual maps.

**Acknowledgments:** We would like to thank Francis Crick, Ellen Hildreth, James Mahoney and Tomaso Poggio for their enlightening comments. Gady Geiger pointed out the difference between theories and experimental data.

## References

Atkinson,J., Campbell,F.W. & Francis,M.R. "The magic number $4 \pm 0$: A new look at visual numerosity judgments", *Perception* **5**, 327–334, 1976.

Barlow,H.B., "Critical limiting factors in the design of the eye and visual cortex", *Proc. Roy. Soc. Lond. B* **212**, 1–35, 1981.

Beck,J., "Perceptual grouping produced by line figures", *Perception & Psychophysics* **2**, 491–495, 1967.

Bergen,J.R. & Julesz,B., "Focal attention in rapid pattern discrimination", *Nature* **303**, 696–698, 1983.

Bushnell,C., Goldberg,M.E. & Robinson,D.L., "Behavioral enhancement of visual responses in monkey cerebral cortex. I. Modulation in posterior parietal cortex related to selective visual attention", *J. Neurophysiol.* **4**, 755–772, 1981.

Crick,F., "The function of the thalamic reticular complex: the searchlight hypothesis", *Proc. Natl. Acad. Sci. USA*, In press, 1984.

Engel,F.L., "Visual conspicuity, directed attention and retinal locus", *Vision Res.* **11**, 563–576, 1971.

Engel,F.L., "Visual conspicuity and selective background interference in eccentric vision", *Vision Res.* **14**, 459–471, 1974.

Eriksen,C.W. & Schultz,D.W., "Retinal locus and acuity in visual information processing", *Bull. Psychonomic Soc.* **9**, 81–84, 1977.

Feldman,J.A.: Dynamic connections in neural networks. *Biol. Cybern.* **46**, 27–39, 1982.

Feldman,J.A. & Ballard,D.H., "Connectionist models and their properties", *Cognitive Science* **6**, 205–254, 1982.

Geiger, G. & Lettvin, J., "On enhancing the perception of forms in peripheral vision". Submitted to *Nature*, 1984.

Gilbert, C.D. & Kelly, J.P., "The projections of cells in different layers of the cat's visual cortex", *J. Comp. Neur.*, **163**, 81–106, 1975.

Graybiel, A.M. & Berson, D.M., "On the relation between transthalamic and transcortical pathways in the visual system", In: *The organization of the cerebral cortex*, Eds. F.O. Schmitt, F.G. Worden, G. Adelman, S.G. Dennis, MIT Press, Cambridge, 1981.

Hadeler, K.P., "On the theory of lateral inhibition", *Kybernetik*, **14**, 161–165, 1974.

Julesz,B., "Textons, the elements of texture perception and their interactions", *Nature* **290**, 91–97, 1981.

Julesz,B., "A brief outline of the texton theory of human vision", *Trends Neurosci.* **7**, 41–48, 1984.

Julesz,B. & Bergen,J.R., "Textons, the fundamental elements in preattentive vision and perception of textures", *Bell System Tech. J.* **62**, 1619–1645, 1983.

Koch,C., Poggio,T. & Torre,V., "Retinal ganglion cells: a functional interpretation of dendritic morphology", *Phil. Trans. R. Soc. Lond. B*, **298**, 227–264, 1982.

Land, E.H. "Experiments in color vision", *Sci. Am.*, **200**, 84–89, 1959.

Llinas, R. & Sugimori, M., "Electrophysiological properties of *in vitro* purkinje cell dendrites in mammalian cerebellar slices", *J. Physiol.*, **305**, 197–213, 1980.

Macchi, G. & Rinvik, E., "Thalamo-telencephalic circuits: a neuroanatomical survey". In: *Handbook of Electroencephalography and Clinical Neurophysiology*, **Vol. 2(A)**, Ed. O. Creutzfeldt, Elsevier, Amsterdam, 1976.

McCulloch, W.S. & Pitts,W., "A logical calculus of the ideas immanent in nervous activity", *Bull. math. Biophys.*, **5**, 115-133, 1943.

Minsky,M., "K–lines: a theory of memory", *Artificial Intelligence Lab. Memo* **No. 516**, MIT Cambridge, 1979.

Minsky,M. & Papert,S., *Perceptrons*, Cambridge, Massachusetts, MIT press, 1969.

Neisser,U., *Cognitive Psychology.* Appleton-Century-Crofts, New York, 1967.

Poggio, T., "Computational geometry and the processing focus: Attention as the router", Unpublished manuscript, 1984.

Posner, M.I., "Orienting of attention", *Quart. J. exp. Psychol.* **32**, 3–25, 1980.

Posner, M.I., Cohen, Y. & Rafal, R.D., "Neural systems control of spatial orienting", *Phil. Trans. R. Soc. Lond. B*, **298**, 187–198, 1982.

Posner, M.I., Snyder,C.R.R. & Davidson,B.J., "Attention and the detection of signals", *J. exp. Psychol.: General* **109**, 160–174, 1980.

Remington, R. & Pierce,L., "Moving attention: evidence for time-invariant shifts of visual selective attention", *Perception & Psychophysics*, **35(4)**, 393–399, 1984.

Rosenfeld, A., *Multiresolution image processing and analysis*, Editor, Springer Verlag, Berlin, 1984.

Schmitt, F.O., Dev, P. & Smith, B.H., "Electrotonic processing of information by brain cells", *Science*, **193**, 114–120, 1976.

Sherman, S.M., "Functional organization of the W-, X-, and Y-cell pathways in the cat: a review and hypothesis", In: *Progress in Psychobiology and Physiological Psychology*, Eds. J.M. Sprague and A.N. Epstein, Academic Press, New York, 1984.

Shulman, G.L., Remington,R.W. & McLean,J.P., "Moving attention through visual space", *J. exp. Psychol.: H.P. & P.* **5**, 522–526, 1979.

Treisman,A., "Focused attention in the perception and retrieval of multidimensional stimuli", *Perception & Psychophysics* **22**, 1–11, 1977.

Treisman,A., "The role of attention in object perception." In *Physical and biological processing of images*, O.J.Braddick and A.C. Sleigh, Editors. Springer Verlag , Berlin, 1983.

Treisman,A. & Gelade,G., "A feature-integration theory of attention", *Cog. Psychol.* **12**, 97–136, 1980.

Tsal,Y., "Movements of attention across the visual field", *J. exp. Psychol.: H.P. & P.* **9**, 523–530, 1983.

Tsumoto, T., Creutzfeldt, O.D. and Legendy, C.R., "Functional organization of the corticofugal system from visual cortex to lateral geniculate nucleus in the cat", *Exp. Brain Res.*, **32**, 345–364, 1978.

Ullman,S., "Visual Routines". *Artificial Intelligence Lab. Memo* **No. 723**, MIT, Cambridge, 1983.

Wertheimer,M., " Untersuchungen zur Lehre von der Gestalt II. Psychol, Forsch.", **4**, 301–350, 1923.

Wilson,H.R. & Bergen,J.R., "A four mechanism model for threshold spatial vision". *Vision Res.* **19**, 19–32, 1979.

Wong, R.K.S., Prince, D.A. & Basbaum, A.I.: "Intradendritic recordings from hippocampal neurons", *Proc. Natl. Acad. Sci. USA*, **76**, 986–990, 1979.

Yingling, C.D. & Skinner, J.E., "Gating of thalamic input to cerebrai cortex by nucleus reticularis thalami". In: *Attention, voluntary contraction and event-related cerebral potentials. Prog. Clin. Neurophysiol.*, **1**, Ed. J.E. Desmedt, Karger, Basel, 1977.